OPEN ACCESS Freely available online

PLoS one

Deciphering Network Community Structure by Surprise

Rodrigo Aldecoa, Ignacio Marín*

Instituto de Biomedicina de Valencia, Consejo Superior de Investigaciones Científicas, Valencia, Spain

Abstract

The analysis of complex networks permeates all sciences, from biology to sociology. A fundamental, unsolved problem is how to characterize the community structure of a network. Here, using both standard and novel benchmarks, we show that maximization of a simple global parameter, which we call Surprise (*S*), leads to a very efficient characterization of the community structure of complex synthetic networks. Particularly, *S* qualitatively outperforms the most commonly used criterion to define communities, Newman and Girvan's modularity (Q). Applying *S* maximization to real networks often provides natural, well-supported partitions, but also sometimes counterintuitive solutions that expose the limitations of our previous knowledge. These results indicate that it is possible to define an effective global criterion for community structure and open new routes for the understanding of complex networks.

Citation: Aldecoa R, Marín I (2011) Deciphering Network Community Structure by Surprise. PLoS ONE 6(9): e24195. doi:10.1371/journal.pone.0024195

Editor: Eshel Ben-Jacob, Tel Aviv University, Israel

Received April 14, 2011; **Accepted** August 4, 2011; **Published** September 1, 2011

Funding: This project was supported by grant BIO2008-05067 (Programa Nacional de Biotecnología; Ministerio de Ciencia e Innovación, Spain). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: imarin@ibv.csic.es

Introduction

A network of interacting units is often the best abstract representation of real-life situations or experimental data. This has led to a growing interest in developing methods for network analysis in scientific fields as diverse as mathematics, physics, sociology and, most especially, biology, both to study organismic (e. g. populational, ecological) and cellular (metabolic, genomic) networks [1–5]. A significant step to understand the properties of a network consists in determining its communities, compact clusters of densely linked, related units. However, the best way to establish the community structure of a network is still disputed. Many strategies have been used (reviewed in [6]), the most popular being the maximization of Newman and Girvan's modularity (Q) [7]. However, Q has the drawback of being affected by a resolution limit: its maximization fails to detect communities smaller than a threshold size that depends on the total size of the network and the pattern of connections [8]. Since this finding, no other global parameters have been proposed to substitute Q. Alternative strategies (searching for local structural determinants, multilevel optimization of Q) have been suggested, but none of them has achieved general acceptance [6].

Some years ago, we suggested determining the community structure of a network by evaluating the distributions of intra- and inter-community links with a cumulative hypergeometric distribution [9]. Accordingly, to find the optimal community structure of a network of symmetrically connected units (undirected graph) is equivalent to maximize the following parameter:

$$S = -\log \sum_{j=p}^{Min(M,n)} \frac{\binom{M}{j}\binom{F-M}{n-j}}{\binom{F}{n}} \qquad (1)$$

Where F is the maximum possible number of links in a network (i. e. $[k^2-k]/2$, being k the number of units), n is the observed number of links, M is the maximum possible number of intracommunity links for a given partition, and p is the total number of intracommunity links actually observed in that partition. The parameter S, which stands for *Surprise*, indeed measures the "surprise" (improbability) of finding by chance a partition with the observed enrichment of intracommunity links in a random graph.

In this work, we show that S has features that make it the parameter of choice for global estimation of community structure. By using standard and novel benchmarks and a set of high-quality algorithms for community detection, we show that maximizing S often provides optimal characterizations of the existing communities. When this method is applied to real networks, we obtained some expected, logical solutions – some of them much better than those provided by Q maximization – but also unexpected partitions that demonstrate the limitations that the usage of inefficient tools has hitherto cast over the field.

Results

Testing the performance of a global parameter to determine community structure requires both a set of efficient algorithms for community detection and a set of standard benchmarks, consisting in synthetic networks of known structure. In this study, six selected algorithms (see Methods) were tested in two types of benchmarks, which will be called LFR and RC throughout the text. LFR (Lancichinetti-Fortunato-Radicchi) benchmarks are characterized by providing networks in which both the degrees of the nodes and the sizes of the communities follow power laws [10]. RC (Relaxed Caveman) benchmarks start with networks in which all the nodes in a community are connected. Then, this structure is relaxed by generating intercommunity links [11]. We further divided LFR and RC benchmarks into "open" and "closed". Open bench-

marks have been commonly used in the past (e.g. [10,12,13]). In them, sets of similar networks with different proportions of intercommunity links are tested. With many intercommunity links, the networks approach randomness. In closed benchmarks, a starting community structure is progressively transformed into a second, final structure which is exactly known.

For each benchmark, we estimated S and Q with the six algorithms. The maximum values of S and Q obtained (S_{max} and Q_{max}) provided the partitions used to compare with the known community structures. As in previous works [10,14,15], Normalized Mutual Information (NMI) was used to measure the congruence between the known and the estimated community structures. However, we also used the Variation of Information (VI) [16] in a particular case.

Open benchmarks

Figures 1a and 1b summarize the results obtained for four standard open LFR benchmarks that differ in number of units and community sizes [10] (see Methods). Figure 1a indicates that selecting the solution with a maximum S value leads to a perfect characterization of the network structure ($NMI_S = 1$) even when that structure is blurred by a large number of inter-community links, generated by increasing the mixing parameter μ up to 0.5–0.7 (see Methods for μ definition). If μ is further increased, the original partition is not chosen by any algorithm ($NMI_S < 1$). This suggests that the original community structure is not present anymore, which is in good agreement with the fact that $S_{max} \gg S_{orig}$, where S_{orig} is the S value obtained assuming that the original community structure is still present (Table S1). S maximization qualitatively improves over Q maximization (Figure 1b and Table S1): $NMI_S > NMI_Q$ in 2827/3600 = 78.5% of the cases, $NMI_Q > NMI_S$ in just 4.1% of them and the rest are ties. Interestingly, $NMI_Q \ll NMI_S$ in quasi-random and random networks (Figure 1b), suggesting that maximizing Q overimposes spurious community structures in those cases. It is significant that S maximization provided better average NMI scores than those obtained by any single algorithm in these same benchmarks [15]. Different algorithms provided the top S scores, depending on the benchmark and μ value examined (Figure 2a and Figure S1).

The discovery of the resolution limit of Q showed that heterogeneous community sizes may greatly affect the ability of global parameters to detect structure [8]. However, by construction, community sizes in the standard LFR benchmarks are very similar. Pielou's evenness indexes (PI) [17] ranged from 0.96 to

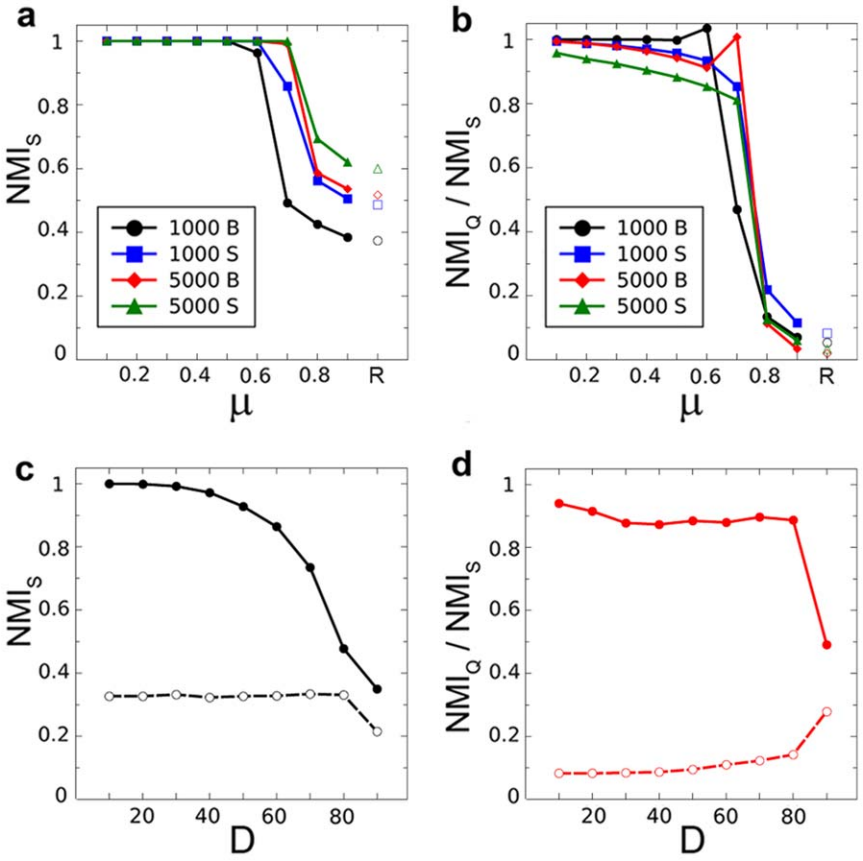

Figure 1. Results for open LFR and RC benchmarks. a) Results for the four standard LFR networks. B and S indicate big and small communities respectively and 1000 or 5000 the number of nodes. μ: mixing parameter. NMI measures the congruence between the known and the deduced community structures. Each point is based on 100 different networks; standard errors of the mean are too small to be visualized. Values for 100 random (R) networks with the same number of units and degree distributions are also shown. b) Comparison of S and Q maximizations in LFR benchmarks. The NMI_Q/NMI_S ratios, which are almost always below 1, are shown. c) Results for the RC benchmark. The parameter Degradation (D) indicates the percentage of both deleted and shuffled links. Each black dot is based on 100 networks, again standard errors are so small that cannot be visualized at this scale. For each value of D, results for 100 random networks with the same number of links are also shown (open circles). d) Relative quality of the partitions generated by maximizing S and Q in RC benchmarks. As in panel b, NMI_Q/NMI_S ratios are shown. White dots: results for random networks with different D values.
doi:10.1371/journal.pone.0024195.g001

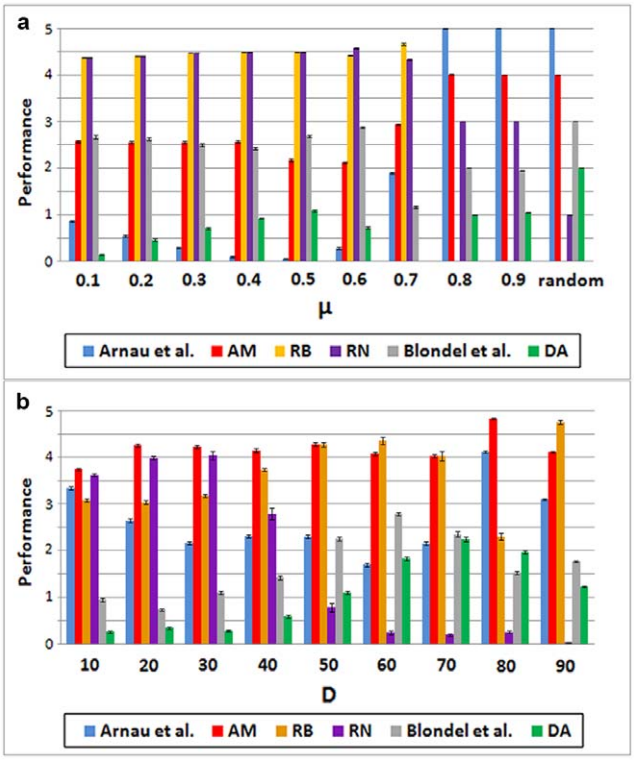

Figure 2. Average performance of the algorithms in the open LFR and RC benchmarks. The algorithms used were described by Arnau *et al.* [9], Aldecoa and Marín (AM) [13], Rosvall and Bergstrom (RB) [23], Ronhovde and Nussinov (RN) [24], Blondel *et al.* [25] and Duch and Arenas (DA) [26]. a) Typical example of the results obtained in LFR benchmarks, here with 5000 units and big communities (see Figure S1 for all of them). After ordering the algorithms from best to worst performance, their ranks were added for the 100 different networks. Performance was defined as P = 6 - average rank. Therefore, the maximum value P = 5 means that an algorithm was the best in all networks tested, while P = 0 means that it was always the worst. As it can be observed, none of the algorithms achieved optimal results in all cases. b) Results obtained in the RC benchmark with different Degradation (D) values. Performance evaluated as in panel a).
doi:10.1371/journal.pone.0024195.g002

0.98 in the four benchmarks used above, close to the maximum value of the index (PI = 1 for communities of identical size). Considering that it was critical to test S in more extreme situations, we built the RC benchmarks, which have PIs as low as 0.70 (as shown in Figure S2). Figures 1c and 1d summarize the results for open RC benchmarks, with progressive Degradation (D; see Methods) of the original structure. That structure is efficiently detected by S maximization, with a slow decrease in performance when D increases (Figure 1c; see also Table S2, Figure S2). Again, S maximization clearly improves over Q maximization in these benchmarks (Figure 1d; $NMI_S > NMI_Q$ in 848/900 = 94.2% of the cases, while $NMI_Q > NMI_S$ in just 3.3% of the cases). As occurred for the LFR benchmarks, none of the algorithms obtained the best results in all networks (Figure 2b).

Closed benchmarks

The results just shown indicate that using S_{max} to detect community structure has obvious advantages over maximizing Q. However, they do not allow to evaluate how optimal is that criterion, given that the potential maximum NMIs are unknown. To solve this limitation, we generated closed LFR and RC benchmarks, in which we had an a priori expectation of the maximum NMI values. Results are shown in Figures 3 (LFR) and 4 (RC). In all cases in which S_{max} was used, an almost perfectly symmetrical dynamics was observed. In the process of converting the original structure into the final one (by increasing the Conversion parameter; see Methods), NMI losses for the first structure are compensated by increases for the second. The average of both NMIs is thus approximately constant, and it has a value identical or very close to $(1+NMI_{IF})/2$, where NMI_{IF} is obtained comparing the initial and final structures (Figures 3a–d; Figures 4a–c; Figures S3, S4). This is exactly the result expected for an optimal parameter (see theoretical details in Methods). On the contrary, maximizing Q shows a poor performance except when community sizes are very similar/identical (Figures 3e, 4d; Figures S3, S4). The same results were obtained using a second measure of congruence, Variation of Information (VI) (Figures S5, S6). Finally, in the LFR benchmarks, S_{max} was always identical or higher than S_{orig} (Figure 3f). However, this does not happen for the RC benchmarks (Figure 4e). Therefore, these algorithms sometimes fail to obtain the highest possible S values. This fact may explain the slight departures from NMI symmetry observed in some RC benchmarks (blue diamonds in Figures 4b, 4c).

Real networks

Figure 5 summarizes the S_{max} results for three real networks. The first example is based on the CYC2008 database, which compiles 1604 proteins that belong to 324 protein complexes [18]. The general agreement between communities detected using S_{max} and a priori defined protein complexes is almost perfect, $NMI_S = 0.91$. On Figure 5a, the 11 communities of size >20, out of the 313 detected, are detailed to show how fine-grained is the classification obtained. On the contrary, optimizing Q provides a very coarse classification into just 24 communities with $NMI_Q = 0.57$. The largest five communities alone almost cover the whole network (Figure 5b). These results indicate how excellent is S performance when there are many small, abundant communities, a typical situation in which Q, affected by its resolution limit, radically fails. Figure 5c shows, as a positive control, the results for a classical benchmark of well-known structure, the *College football network* [12]. The agreement with the expected communities is again very high ($NMI_S = 0.93$). Finally, Figure 5d shows the results for another well-known example, the *Zachary's Karate club* network [12,19]. This social network supposedly contains two communities. However, S analyses surprisingly unearthed 19 communities, 12 of them singletons (Figure 5d).

Discussion

In this study, we have shown the potential of maximizing the global parameter Surprise (S) to determine the community structure present in complex networks. The results indicate that it has a qualitative better performance than the hitherto most commonly used global measure, Newman and Girvan's modularity (Q). The advantage of S over Q is maybe not that surprising, considering the different theoretical foundations of both measures. Newman and Girvan's Q is based on a simple definition of community, as a region of the network with an unexpectedly high density of links. However, the number of units within each community does not influence the value of Q [7]. On the contrary, S evaluates both the number of links and of units in each community (see Formula (1)). Therefore, S implicitly assumes a more complex definition of community: a precise number of units for which it is found a density of links which is statistically unexpected given the features of the network. In this context of comparison of both measures, it is also very significant that, while

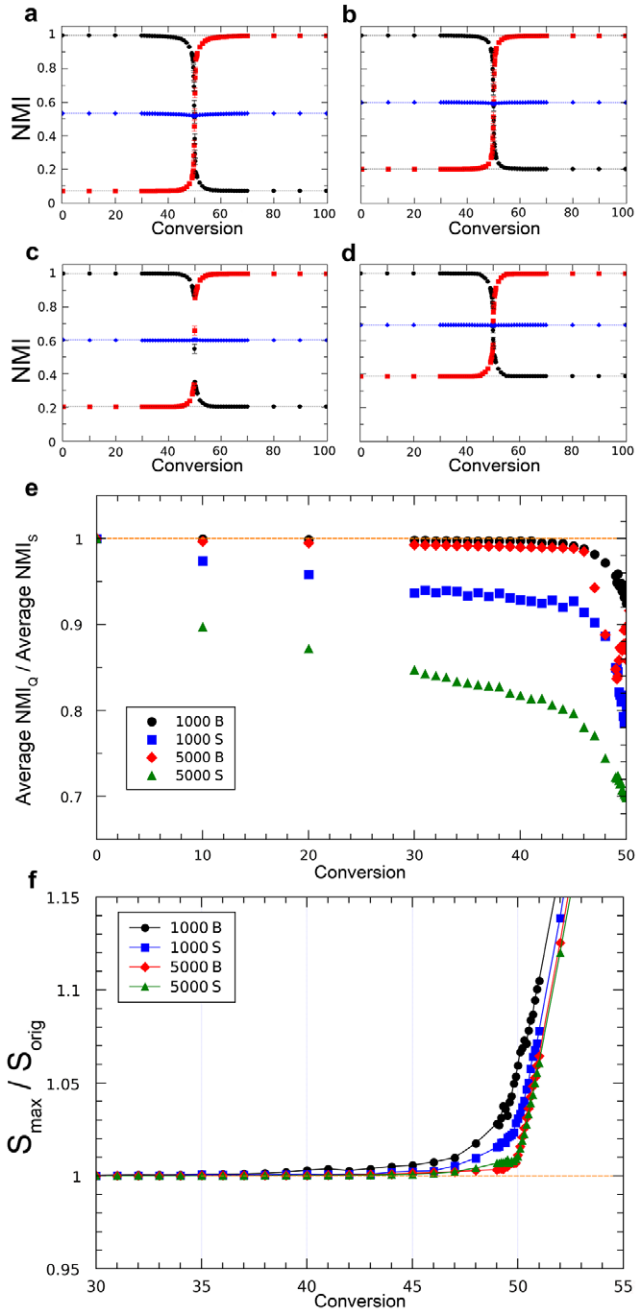

Figure 3. Results for closed LFR benchmarks. a) LFR benchmark with 1000 units and big communities. For each Conversion (C) value, NMIs comparing the S_{max} partition with the initial (black dots) or final (red squares) community structures were obtained. The symmetrical results led to NMI averages (blue diamonds) that, with great precision, fell in a straight line of value $(1+NMI_{IF})/2$. Dots are based on 100 independent analyses. b–d) LFR benchmarks with, respectively, 1000 units, small communities (b), 5000 units, big communities (c) and 5000 units, small communities (d). Results are very similar to those in panel a). e) Average NMI values for partitions obtained maximizing Q are worse than those obtained maximizing *S*, especially as we move towards C = 50, in which the real community structure is more difficult to establish. This effect is exacerbated by large number of units and small community sizes, due to the resolution limit of Q. Results for C>50 are symmetrical to the ones shown here. See also Figure S3. f) S_{max}/S_{orig} ratio ≥1, i. e. either the original structure or a different one with higher *S* is found. These results are compatible with the algorithms used being able to detect the true structure present with great accuracy.
doi:10.1371/journal.pone.0024195.g003

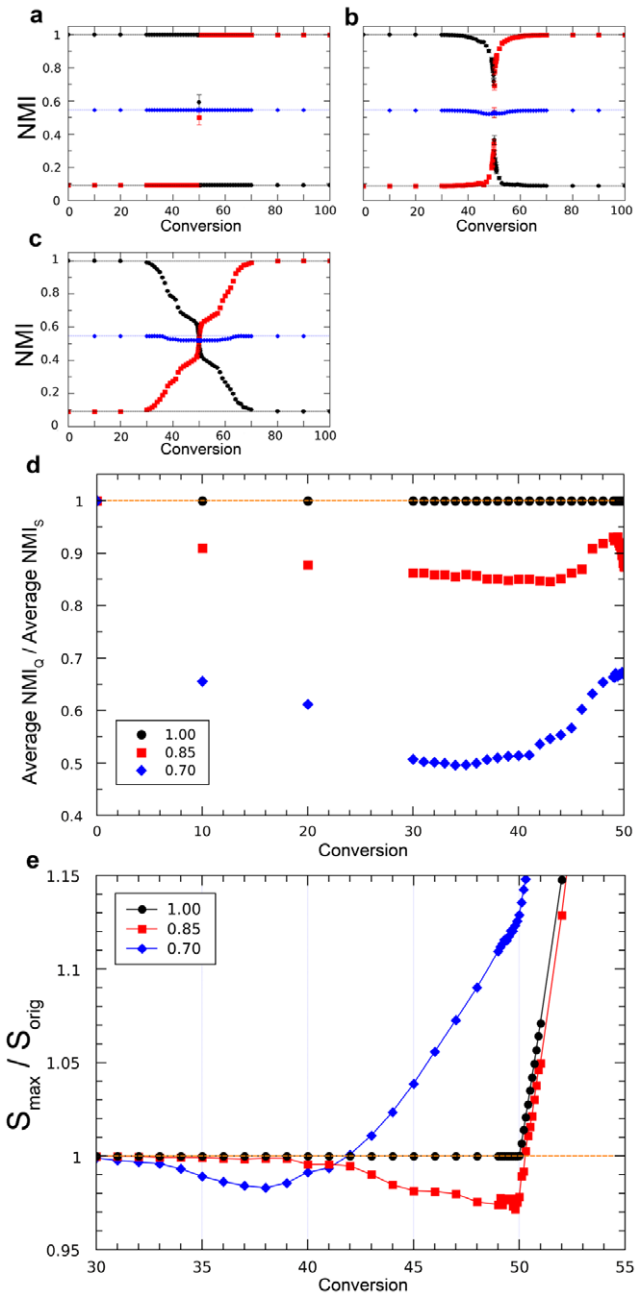

Figure 4. Results for closed RC benchmarks. Three networks with different heterogeneity in community sizes (Pielou's indexes equal to 0.70, 0.85 and 1.00 respectively) were used as examples. a) PI = 1; b) PI = 0.85; c) PI = 0.70. Results similar to those in Figure 2, except that the figures are not so perfectly symmetrical in the most heterogeneous networks (panels b and c; blue diamonds slightly deviate from the straight line). d) Average NMI values are much worse when Q is used, provided that community sizes are heterogeneous. See also Figure S4. e) $S_{max}/S_{orig}<1$ with heterogeneous community sizes. The algorithms used did not detect in those cases the maximum possible *S*, which still may correspond to the initial structure. This may contribute to the departures from symmetry shown in panel a). The fact that $S_{max}/S_{orig} \gg 1$ with C<0.50 and PI = 0.70 (blue diamonds) implies that the algorithms are detecting structures different from the initial one.
doi:10.1371/journal.pone.0024195.g004

some of the algorithms used in this work were the best among those specifically designed to maximize Q, none was devised to maximize *S*. Therefore, our results actually underestimate the

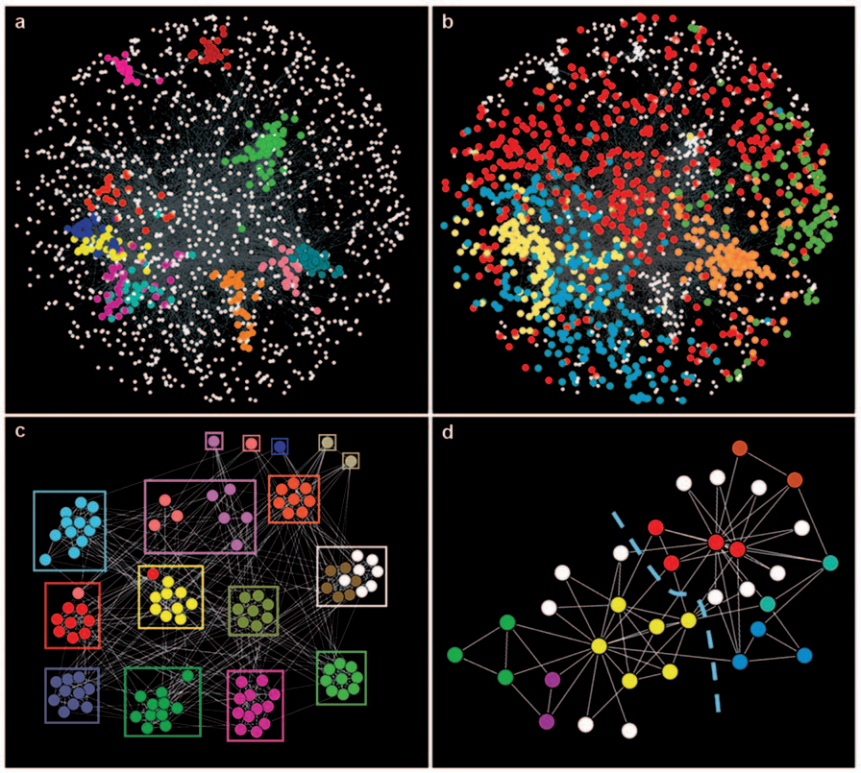

Figure 5. S_{max} analyses applied to real networks. Community structure of the CYC2008 network (a, b), College football network (c) and Zachary's karate club network (d), according to S maximization (panels a, c, d) or Q maximization (panel b). In panel c, the known community structure is shown (squares). The broken lines in panel d divide the network into the two communities assumed to exist. That division of the network is not supported at all by S_{max} analyses. While $S_{(2\ communities)} = 13.61$, the optimal division found has $S_{(19\ communities)} = 25.69$. Twelve of these optimal communities are singletons (white dots).
doi:10.1371/journal.pone.0024195.g005

power of S maximization for community detection. A direct example of that underestimation is shown in Figure 4e: the maximum values of S were, in some cases, not found. The few exceptions found in which $NMI_Q > NMI_S$ (3–4% of all the cases examined in the open benchmarks) could be also explained by an incomplete success in determining S_{max} with these algorithms.

The commonly used open benchmarks are useful for general evaluations of the performance of different algorithms, but they do not allow to establish how optimal are the results obtained. For that, we have devised novel closed benchmarks in which an initial known community structure is progressively transformed into a second, also known, community structure. Provided that both community structures are identical, it can be demonstrated that, at any point of the transformation from one to the other, the average of the NMIs of the solution found respect to the initial and final structures should approximate a constant value ($[1+NMI_{IF}]/2$), if that solution is optimal (see Methods). This feature allows establishing the intrinsic quality of the partitions obtained, with S maximization often providing optimal results. We conclude that S maximization establishes the community structure of complex networks with a high accuracy. Two promising lines of research are clear. First, generating novel, specific algorithms for S maximization, which may improve over the existing ones. Second, building a standard set of closed benchmarks to test any new algorithms for community detection. Our LFR and RC closed benchmarks may be a good starting point for that standard set.

When S maximization was applied to real networks, the results obtained are of two types. On one hand, for the CYC2008 and College football networks, the expectation was to find a clear community structure which should faithfully correspond to either the complexes to which the proteins examined are part (CYC2008 network) or to the conferences to which the teams belong (College football network), given that intracomplex or intraconference links are abundant (e. g. Figure 5c). These are exactly the results found using S_{max}. On the other hand, the structure of the Zachary's karate network is far from obvious (Figure 5d). Therefore, finding that, according to S_{max}, the network contains some small groups plus many singletons is, at least a posteriori, not so unexpected. A natural question is then why the scientific community has been so keen of exploring this particular network, often to establish whether an algorithm was able or not to detect the putative two communities [e. g. refs. 7, 12, 19, 20 among many others]. This may reflect a psychological bias, to which the use of under-performing methods for community detection may have certainly contributed. It shows to which extent human prejudices may taint evaluations in this type of ill-defined problems.

Methods

Algorithms used to maximize S and Q

Six of the best available algorithms, selected either by their exceptional performance in artificial benchmarks or their success in previous analyses of real and simulated networks [9,13–15,21,22], were used. They were the following: 1) UVCluster algorithm [9,13]: It performs iterative hierarchical clustering, generating dendrograms. The best values of S and Q were obtained scanning these dendrograms from root to leaves. 2) SCluster algorithm [13]: also performs iterative hierarchical

clustering, but using an alternative strategy which is faster and sometimes more accurate than the one implemented in UVCluster. 3) Dynamic algorithm by Rosvall and Bergstrom [23]: an algorithm based on expressing the characterization of communities as an information compression problem. 4) Potts model multiresolution algorithm [24]: works by minimizing the Hamiltonian of a Potts spin model at different resolution scales, i. e. searching for communities of different sizes. 5) Fast modularity optimization [25]: devised to maximize Q. It provides multiple solutions from which values for *S* and Q can be obtained, and the maximum ones were used in our analyses. 6) Extremal optimization algorithm [26]: A divisive algorithm also developed to maximize Q. Analyses were always performed with the default program settings.

Features of the benchmarks

First, the recently developed LFR benchmarks, specifically devised for testing alternative community detection strategies [10], were used. In particular, we chose four standard LFR benchmarks already explored by other authors [15]. The networks analyzed had either 1000 or 5000 units and were built according to two alternative ranges of community sizes (Big (B): 20–100 units/community; Small (S): 10–50 units/community). For each of the four conditions (1000 B, 1000 S, 5000 B, 5000 S), 100 different networks were generated for each value of a mixing parameter μ, which varied from 0.1 to 0.9 [15]. μ is the average percentage of links that connect a unit to those in other communities. Logically, increasing μ weakens the network community structure. When $\mu = 0.9$, the networks are quasi-random (see below).

Once found that these LFR benchmarks generated networks with communities of very similar sizes, we decided to implement RC benchmarks in which these sizes were more variable. All networks in these benchmarks had 512 units divided into 16 communities. One hundred networks with random community sizes, determined using a broken-stick model [27], were generated. This model provides highly heterogeneous community sizes. Progressive weakening of the community structure of the RC networks, similar to the effect of increasing μ in the LFR networks, was obtained as follows. Initially, all units of each community in the network were fully connected. Then, that obvious structure was progressively blurred, by first randomly removing a certain percentage of edges and then randomly shuffling the same percentage of links among the units. That common percentage, we have called Degradation (D). Thus, D = 10% means that, first, 10% of the links present were eliminated and then 10% of the remaining edges were randomly shuffled among units. Shuffling involved first the random removal of an edge of the graph and then the addition of a new edge between two randomly chosen nodes.

In the LFR and RC benchmarks just described it was possible to compare networks having obvious community structures (generated with low μ or D parameters) with others that were increasingly random. This type of benchmarks, we have called open. We also generated closed LFR and RC benchmarks. In them, links were shifted in a directed way, in order to convert the original community structure of a network into a second, also predefined, structure. In this way, it is possible to monitor when the original structure is substituted by the final one according to the solutions provided by S_{max} or Q_{max}. In the LFR and RC closed benchmarks, the starting networks were the same described in the previous paragraphs, with $\mu = 0.1$ (LFR) or D = 0 (RC) respectively, and the final networks were obtained by randomly relabeling the nodes. Therefore, the initial and final networks had identical community structures but the nodes within each community were different. Conversion (C) is defined as the percentage of links exclusively present in the initial network that are substituted by links only present in the final one (i. e. C = 0: initial structure present; C = 100: final structure present).

NMI symmetry as a measure of performance in closed benchmarks

In our closed benchmarks, a peculiar symmetrical behavior of NMI values respect to the initial and final partitions is expected. Imagine that a putative optimal partition is estimated according to a given criterion. Let us now consider the following triangle inequality:

$$(\mathrm{NMI_{IE}} + \mathrm{NMI_{EF}})/2 \leq (1 + \mathrm{NMI_{IF}})/2 \qquad (2)$$

where $\mathrm{NMI_{IE}}$ is the normalized mutual information calculated for the initial structure (I) and the estimated partition (E), $\mathrm{NMI_{EF}}$ is the normalized mutual information for the final structure (F) versus the estimated partition and $\mathrm{NMI_{IF}}$ is the normalized mutual information for the comparison between the initial and final structures. Inequality (2) holds true if the structures of I, F and E are identical (i. e. both the number and sizes of the communities are the same, but not necessarily are the same the nodes within each community). This follows from the fact that

$$1 - \mathrm{NMI_{XY}} = \mathrm{VI_{XY}}/[(\mathrm{H}(\mathrm{X}) + \mathrm{H}(\mathrm{Y})] \qquad (3)$$

Where $\mathrm{VI_{XY}}$ is the Variation of Information for both partitions [16] and H(X) and H(Y) are the entropies of the X and Y partitions, respectively. Given that VI is a metric [16], it satisfies the triangle inequality

$$\mathrm{VI_{AB}} + \mathrm{VI_{BC}} \geq \mathrm{VI_{AC}} \qquad (4)$$

If, as indicated, the structures of all partitions are identical, then all their entropies are also identical. In that case, the following inequality can be deduced from formulae (3) and (4):

$$(1 - \mathrm{NMI_{AB}}) + (1 - \mathrm{NMI_{BC}}) \geq (1 - \mathrm{NMI_{AC}}) \qquad (5)$$

From this inequality, and substituting A, B and C with I, E and F, respectively, formula (2) can be deduced. Formula (2) therefore means that, provided that I, E and F have the same structure, the average of $\mathrm{NMI_{IE}}$ and $\mathrm{NMI_{EF}}$ may acquire a maximum value $[(1+\mathrm{NMI_{IF}})/2]$. Inequality (2) will also hold approximately true if the entropies of I, E and F are very similar (i. e. many identical communities). In our closed benchmarks the I and F structures are identical, and we progressively convert one into the other. It is thus expected that the optimal partition along this conversion is similar in structure to both I and F. Hence, deviations from the expected average value $(1+\mathrm{NMI_{IF}})/2$ are a cause of concern, as they probably mean that the optimal partition has not been found. On the other hand, finding values equal to $(1+\mathrm{NMI_{IF}})/2$ is a strong indication that the optimal partition has indeed been found.

It is worth noting that, although NMI has been commonly used in this field [10,14,15], using VI instead has clear advantages to analyze closed benchmarks: Formula [4] can be used instead of Formula (2), avoiding considering entropies at all. This is why we evaluated the closed benchmark results both using NMI and VI (see above).

Real networks

Two of the three networks explored, known as *College football* and *Zachary's karate* networks, have been frequently used in the past in the context of community detection [e. g. refs. 7, 12, 19, 20, 28]. The third network derived from the CYC2008 protein complexes database [18]. This database contains information for 408 protein complexes of the yeast *Saccharomyces cerevisiae*. The protein complex data were converted into 324 non-overlapping complexes by assigning each protein present in multiple complexes to the largest one. This was made to allow for NMI calculations. Once each protein (unit) was assigned to a non-overlapping cluster (community), we downloaded from the BioGRID database [29] the protein-protein interactions (edges) characterized so far for all these proteins. The final graph contained 1604 nodes and 14171 edges.

Supporting Information

Figure S1 Average performances of the algorithms in the LFR benchmarks. With different network sizes (1000, 5000 units), community sizes (small: 10 to 50 units per community; big: 20–100 units per community) and values of mixing parameter (μ) and for random networks of the same size. After ordering the algorithms from best to worst performance, their ranges were added for the 100 different networks. Performance is defined as P = 6 - average range.
(TIF)

Figure S2 Details of the results for the RC benchmark. a) Normalized Mutual Information values for the 100 networks tested, obtained by *S* maximization. Given that both a low Pielou's index and high D may alter the original structure of the network, these results would tend to underestimate the real quality of the partition into communities obtained. Lines correspond to the second degree polynomials that best fit the results, which were found to be better than the first degree ones. b) Examples of the relative sizes of communities for different Pielou's indexes, to show the very different structures provided by generating the community sizes according to a broken stick model. c) Summary of the results in the RC benchmark with Q maximization. The results are much worse than those shown in panel a), due to the resolution limit that affects Q values when some communities are small (low Pielou's indexes). Lines again correspond to the best fits according to second degree polynomials.
(TIF)

Figure S3 Behavior of *S* and Q maximization in closed LFR benchmarks. Notice the obvious decrease below $(1+NMI_{IF})/2$ when Q is maximized.
(TIF)

Figure S4 Results for *S* and Q maximization in the closed RC benchmarks. The behavior of S_{max} is again qualitatively better than the one of Q_{max}, except when all communities are identical.
(TIF)

Figure S5 Behavior of *S* and Q maximization in closed LFR benchmarks using Variation of Information (VI) as a measure of congruence. As it can be deduced from Formula [4] in the main text, a good behavior of a parameter implies minimal deviations from the expected value $VI_{IF}/2$ (blue line). Results are almost identical to those shown in Figure S3 using NMI. S_{max} behavior is clearly better than Q_{max} behavior.
(TIF)

Figure S6 Results for *S* and Q maximization in the closed RC benchmarks, measured with VI. The behavior of S_{max} is again qualitatively better than the one of Q_{max}, confirming the results shown in Figure S5.
(TIF)

Table S1 Detailed results obtained for the LFR benchmarks. The values of NMI when *S* and Q are maximized are indicated, together with the percentage of cases in which NMI = 1 and the values of S_{max} and S_{orig} (i. e. the *S* value obtained assuming that the original structure is present). Notice that when μ = 0-6-0.7, $S_{max} > S_{orig}$, meaning that the original structure is not the one present anymore. In those cases, NMIs are expected to rapidly decrease, as indeed is observed.
(DOC)

Table S2 Details of the RC benchmark results. Same data as in Table S1, but with variations in the Degradation (D) parameter. Data for random networks of the same size are also included.
(DOC)

Author Contributions

Conceived and designed the experiments: RA IM. Performed the experiments: RA. Analyzed the data: RA. Contributed reagents/materials/analysis tools: RA. Wrote the paper: IM.

References

1. Barabási AL, Oltvai ZN (2004) Network biology: Understanding the cell's functional organization. Nat Rev Genet 5: 101–113.
2. Wasserman S, Faust K (1994) Social network analysis: Methods and applications. Cambridge University Press, Cambridge, U.K.
3. Strogatz SH (2001) Exploring complex networks. Nature 410: 268–276.
4. Costa LD, Rodrigues FA, Travieso G, Boas PRV (2007) Characterization of complex networks: A survey of measurements. Adv Phys 56: 167–242.
5. Newman MEJ (2010) Networks: An introduction. Oxford University Press, Oxford. U.K.
6. Fortunato S (2010) Community detection in graphs. Phys Rep 486: 75–174.
7. Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69: 026113.
8. Fortunato S, Barthélemy M (2007) Resolution limit in community detection. Proc Natl Acad Sci USA 104: 36–41.
9. Arnau V, Mars S, Marín I (2005) Iterative cluster analysis of protein interaction data. Bioinformatics 21: 364–378.
10. Lancichinetti A, Fortunato S, Radicchi F (2008) Benchmark graphs for testing community detection algorithms. Phys Rev E 78: 046110.
11. Watts DJ (1999) in Small worlds. The dynamics of networks between order and randomness. Princeton University Press, Princeton, N.J.
12. Girvan M, Newman MEJ (2002) Community structure in social and biological networks. Proc Natl Acad Sci USA 99: 7821–7826.
13. Aldecoa R, Marín I (2010) Jerarca: Efficient analysis of complex networks using hierarchical clustering. PLoS ONE 5: e11585.
14. Danon L, Díaz-Guilera A, Duch J, Arenas A (2005) Comparing community structure identification. J Stat Mech P09008.
15. Lancichinetti A, Fortunato S (2009) Community detection algorithms: a comparative analysis. Phys Rev E 80: 056117.
16. Meilă M (2007) Comparing clusterings – an information based distance. J Multivar Anal 98: 873–895.
17. Pielou EC (1966) The measurement of diversity in different types of biological collections. J Theor Biol 13: 131–144.
18. Pu SY, Wong J, Turner B, Cho E, Wodak SJ (2009) Up-to-date catalogues of yeast protein complexes. Nucl Acids Res 37: 825–831.
19. Zachary WW (1977) Information-flow model for conflict and fission in small-groups. J Anthropol Res 33: 452–473.
20. Freeman LC (1993) Finding groups with a simple genetic algorithm. J Math Sociol 17: 227–241.
21. Lucas JI, Arnau V, Marín I (2006) Comparative genomics and protein domain graph analyses link ubiquitination and RNA metabolism. J Mol Biol 357: 9–17.
22. Marco A, Marín I (2009) Interactome and Gene Ontology provide congruent yet subtly different views of a eukaryotic cell. BMC Syst Biol 3: 69.
23. Rosvall M, Bergstrom CT (2008) Maps of random walks on complex networks reveal community structure. Proc Natl Acad Sci USA 105: 1118–1123.

24. Ronhovde P, Nussinov Z (2009) Multiresolution community detection for megascale networks by information-based replica correlations. Phys Rev E 80: 016109.
25. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech P10008.
26. Duch J, Arenas A (2005) Community detection in complex networks using Extremal Optimization. Phys Rev E 72: 027104.
27. MacArthur RH (1957) On the relative abundance of bird species. Proc Nat Acad Sci USA 43: 293–295.
28. Newman MEJ (2006) Modularity and community structure in networks. Proc Natl Acad Sci USA 103: 8577–8582.
29. Breitkreutz BJ, Stark C, Reguly T, Boucher L, Breitkreutz A, et al. (2008) The BioGRID interaction database: 2008 update. Nucl Acids Res 36: D637–D640.